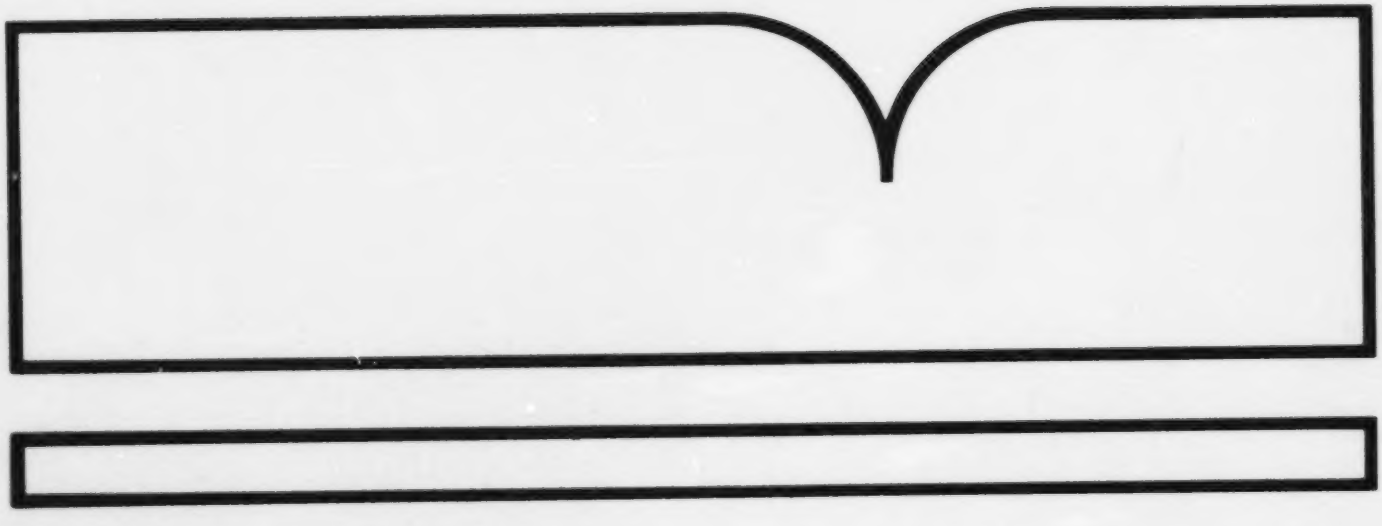
PB86-196706

Applications of Information Theory and Pattern Recognition to Gas Chromatography-Mass Spectrometric Analysis of Toxic Organic Compounds in Ambient Air

(U.S.) Environmental Monitoring Systems Lab.
Research Triangle Park, NC

May 86



U.S. Department of Commerce
National Technical Information Service
NTIS

PB86-196706

EPA/600/D-86/103
May 1986

APPLICATIONS OF INFORMATION THEORY AND PATTERN RECOGNITION TO GAS CHROMATOGRAPHY-MASS SPECTROMETRIC ANALYSIS OF TOXIC ORGANIC COMPOUNDS IN AMBIENT AIR

by

Donald R. Scott
Environmental Monitoring Systems Laboratory
U.S. Environmental Protection Agency
Research Triangle Park, NC 27711

ENVIRONMENTAL MONITORING SYSTEMS LABORATORY
OFFICE OF RESEARCH AND DEVELOPMENT
U.S. ENVIRONMENTAL PROTECTION AGENCY
RESEARCH TRIANGLE PARK, NC 27711

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA/600/D-86/103	2.
3. RECIPIENT'S ACCESSION NO. PB86 196706/AS	
4. TITLE AND SUBTITLE: Applications of Information Theory and Pattern Recognition to Gas Chromatography-Mass Spectrometric Analysis of Toxic Organic Compounds in Ambient Air	5. REPORT DATE: May 1986
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S): Donald R. Scott	8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS: U. S. Environmental Protection Agency, Environmental Monitoring Systems Laboratory, Methods Development and Analysis Division, Research Triangle Park, N. C. 27711	10. PROGRAM ELEMENT NO.
	11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS: ENVIRONMENTAL MONITORING SYSTEMS LABORATORY, OFFICE OF RESEARCH AND DEVELOPMENT, U.S. ENVIRONMENTAL PROTECTION AGENCY, RESEARCH TRIANGLE PARK, NC 27711	13. TYPE OF REPORT AND PERIOD COVERED
	14. SPONSORING AGENCY CODE: EPA/600/08
15. SUPPLEMENTARY NOTES	

16. ABSTRACT

The number of information channels contained in the gas chromatographic, mass spectrometric, and combined gas chromatography-mass spectrometric analysis of 78 toxic organic compounds has been determined. The toxic compounds are those routinely monitored in ambient air samples using Tenax collection and gas chromatography-mass spectrometric analysis. The Shannon information content of the binary encoded and full intensity mass spectra, of the gas chromatographic retention times, and of the combined gas chromatographic-mass spectrometric spectra of the 78 compounds has been calculated. The maximum binary information contents of the 35 channel gas chromatographic, 17 key channel mass spectral, and the 595 channel gas chromatographic-mass spectral methods were 6.4, 15,4, and <21.8 bits, respectively. The 17 masses with the highest binary information content with regard to the 78 compounds were used with SIMCA pattern recognition to determine four classes among the 78 compounds. These included aromatics without chlorine substitution, chloroaromatics, bromoalkanes and alkenes, and chloroalkanes and alkenes. Alkenes and alkanes with both chloro- and bromo substitution were classified as bromo compounds. The principal component models generally consisted of only one component per class, with five masses per class. However, the total alkene and alkane class had two components with twelve masses. Classification accuracy was 96% for the total aromatics and total alkanes and alkenes and 82% for the four subclasses. The pattern recognition study was performed on a commercially available microcomputer with a 64 k CPU.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group

18. DISTRIBUTION STATEMENT: RELEASE TO PUBLIC	19. SECURITY CLASS (This Report): UNCLASSIFIED	21. NO. OF PAGES: 16
	20. SECURITY CLASS (This page): UNCLASSIFIED	22. PRICE

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE

NOTICE

This document has been reviewed in accordance with U.S. Environmental Protection Agency policy and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

APPLICATIONS OF INFORMATION THEORY AND PATTERN RECOGNITION TO GAS CHROMATOGRAPHY-MASS SPECTROMETRIC ANALYSIS OF TOXIC ORGANIC COMPOUNDS IN AMBIENT AIR

Donald R. Scott
Environmental Monitoring Systems Laboratory
U. S. Environmental Protection Agency
Research Triangle Park, North Carolina 27711

The number of information channels contained in the gas chromatographic, mass spectrometric, and combined gas chromatography-mass spectrometric analysis of 78 toxic organic compounds has been determined. The toxic compounds are those routinely monitored in ambient air samples using Tenax collection and gas chromatography-mass spectrometric analysis. The Shannon information content of the binary encoded and full intensity mass spectra, of the gas chromatographic retention times, and of the combined gas chromatographic-mass spectrometric spectra of the 78 compounds has been calculated. The maximum binary information contents of the 35 channel gas chromatographic, 17 key channel mass spectral, and the 595 channel gas chromatographic-mass spectral methods were 6.4, 15.4, and <21.8 bits, respectively. The 17 masses with the highest binary information content with regard to the 78 compounds were used with SIMCA pattern recognition to determine four classes among the 78 compounds. These included aromatics without chlorine substitution, chloroaromatics, bromoalkanes and alkenes, and chloroalkanes and alkenes. Alkenes and alkanes with both chloro- and bromo substitution were classified as bromo compounds. The principal component models generally consisted of only one component per class, with five masses per class. However, the total alkene and alkane class had two components with twelve masses. Classification accuracy was 96% for the total aromatics and total alkanes and alkenes and 82% for the four subclasses. The pattern recognition study was performed on a commercially available microcomputer with a 64 k CPU. The information content of gas chromatography, mass spectrometry, and gas chromatography-mass spectrometry in general and with regard to this analytical problem is discussed.

APPLICATIONS OF INFORMATION THEORY AND PATTERN RECOGNITION TO GAS CHROMATOGRAPHY-MASS SPECTROMETRIC ANALYSIS OF TOXIC ORGANIC COMPOUNDS IN AMBIENT AIR

INTRODUCTION

At the present time the most commonly used analytical procedure for the determination of volatile organic compounds in ambient air is gas chromatography-mass spectrometry. This is due to its high selectivity, sensitivity, and applicability to a very large number of compounds. The present study is concerned with the development of methods for efficient extraction of information regarding chemical class identification from data obtained during routine gas chromatographic-mass spectrometric analyses. The pattern recognition techniques are empirical in nature and therefore do not require any expert knowledge regarding the mass spectra of the classes. The development of such methods would allow retrospective analysis of routine data files for determining the presence or absence of defined chemical classes of interest. The determination of the classes could be done on a small microcomputer without additional analytical costs and does not have to be confined to the compounds originally sought. Such information may be of interest in survey studies of health hazards and in preliminary scans of mass spectral data files before detailed interpretation is accomplished.

A set of 78 toxic organic compounds was chosen from those ca. 100 compounds presently routinely sought in ambient air samples. The set of 78 toxic compounds investigated contained primarily aromatic compounds, haloalkanes, and haloalkenes together with an additional four ethers and epoxides. Approximately 80% of these compounds contain chloro and/or bromo groups. All alkanes and alkenes contained at least one halogen. Shannon information theory and SIMCA pattern recognition were used. The Shannon information contents of the gas chromatographic, mass spectrometric, and the combined gas chromatography-mass spectrometric methods for the identification of the set of 78 compounds also were determined.

THEORETICAL BACKGROUND

Shannon Information Theory

The Shannon information content of a message (1) is related to the reduction in uncertainty gained by the receiver of the message. The information content will depend on the probability of occurrence of the symbols used, correlation between the symbols, encoding and decoding errors, and other uncertainties due to random errors. After chemical analysis the uncertainty with regard to the presence of certain species in the sample is reduced. Therefore the analysis produces a certain amount of Shannon information concerning the sample. In low resolution mass spectrometry the message is equated to the mass spectrum itself. The symbols are the intensities at a given unit mass channel. Neglecting errors, the information content per mass channel, I(j), is given by

$$I(j) = -\sum_{i=1}^{m} p_j(i) \log p_j(i)$$

where m is the number of discrete intensity values available for mass channel j. The probabilities of occurrence of a mass intensity, $p_j(i)$, are calculated over the entire set of reference spectra. The total information

content of a given mass spectrum is the sum of the information contents of the individual channels, if correlation between masses is neglected. Since correlation between masses is well known in mass spectra, this information content should be regarded as an upper limit to the true information content. This formula can also be used to calculate the informational contents of gas chromatography channels (retention time data) and of gas chromatography-mass spectral channels if the distribution of the data representations over the respective channels and compounds are known.

In the case of binary data the Shannon information content can be calculated from a simplified formula. For example with binary encoded spectra there are only two possibilities for the intensity of a mass peak, 1 or 0, corresponding to the presence or absence of the peak above a threshold level. Therefore the information content for mass channel j can be calculated from

$$I(j) = -p_j \log p_j - (1-p_j)\log (1-p_j).$$

Here p_j is the probability of a peak occurring at the mass channel j calculated over the set of mass spectra of the 78 compounds. In the case of the gas chromatography data the binary data corresponds to the presence or absence of a peak in a given channel. In the combined gas chromatography-mass spectral channels the binary data corresponds to the presence or absence of a compound-mass spectral peak in that channel. If the logarithm used is to the base two, then the information content is expressed in bits. The maximum information content per channel for the binary case is one bit and occurs at a probability of 0.5. Nonoccupied channels and those with probabilities of 1.0 contribute no information content.

SIMCA Pattern Recognition

The SIMCA (Soft Independent Modeling of Class Analogy) pattern recognition techniques were developed by Wold and coworkers and have been described in the literature (2,3). The statistical pattern recognition techniques are based on disjoint principal component models for classification of objects. The SIMCA class models are bilinear projection models obtained by decomposing the class data matrix (X) into a score matrix (T) (n x F), a loading matrix (P) (F x p) and a residual matrix (E) (2).

$$(X) = 1 \cdot \bar{x} + (T)\,(P) + (E) \qquad (1)$$

The row vector $\bar{x}$ is composed of all the averages of the variables in the class data matrix. The n x F score matrix (T) describes the projection of the n object points down on the F dimensional hyperplane defined by the F x p loading matrix (P). The residual matrix (E) contains that part of the data matrix due to measurement and modeling errors. If the residuals in (E) are small compared with the variation in (X), then the model is a good representation of (X). When F is two or three, the columns in the score matrix (T) can be plotted against each other to get two dimensional pictures of the objects in hyperspace (measurement space). The modeling power, i.e. the reduction of variance, for a given variable can be used as a measure of its relevance to the class model. Judicious use of the variable loadings and modeling powers obtained in preliminary analyses of the data allows one to polish the model and remove variables which are not significant to the class.

The number of principal components determined and retained for a particular class model is an important consideration. Within the SIMCA

procedures the process of cross validation (4) is used to determine the number of statistically significant principal components for a given class. In this process subsets of the training set of objects are used to fit a class model with one principal component. The objects not used in determining this class model are then fitted to the model, and the sum of squared residuals for the withheld objects is calculated. This is repeated until all of the objects in the training set have been withheld from the model fitting. The overall sum of squared residuals is then calculated for all the withheld objects for this particular model. The entire process is repeated with a class model containing an additional principal component. Addition of principal components to the model is continued until comparison of the sums of squared residuals for the previous and present model show no improvement. Generally, if the number of principal components for a given class model is much smaller than the number of either objects or variables for the class; then the model will be statistically stable.

Once the class models have been determined, objects are classified by fitting their data to the various class models. A standard deviation for each model is calculated from the residuals. This represents a class tolerance level around the principal component model in measurement space. The standard deviations for the objects are calculated from the residuals, and the objects are classified based upon their distances from the class models.

EXPERIMENTAL

Data Set

The low resolution mass spectra of the 78 compounds were obtained from the EPA-NIH Mass Spectral Library on an INCOS data system. A typical spectrum contained approximately 16 peaks. The range of mass/charge ratios was from 35 to 256 with 151 different peaks occurring in the total set. For the full intensity data the intensities were scaled to give a maximum of one for the base peak. The data were also binary encoded by assigning an intensity of one to any peak over the threshold level, which was 1% of the base peak intensity. A list of individual compounds in the set is given in Table I. The retention times were determined on a 50 meter, open tubular, SE 30 wall coated column relative to perfluorotoluene and 1-fluoro -2-iodobenzene internal standards.

Hardware and Software

In this study an Osborne 1 microcomputer (Z80A) with a CP/M operating system and 64 k memory was used. This amount of memory is sufficient to handle a data matrix of size 50 objects by 50 variables. The SIMCA 3B software package was obtained from Principal Data Components, Columbia, Mo.

Data Analysis

The information content of the 151 different nonzero intensity mass channels was calculated from the distribution of the binary encoded mass spectra of the set of 78 compounds using the formulas given above. The full intensity mass spectra of the data set were encoded into ten discrete levels above the threshold level, and the information content for the mass channels also was calculated using this distribution. No correction was made to account for encoding errors or for correlation between mass channels. In the case of the gas chromatography analysis, the width of a channel was determined from the reproducibility (three standard deviations = ca. 0.35 min.) of the relative retention times. The total number of channels was

calculated from the channel width and the time spread from the first and last compounds eluted. The binary information content of the retention times was calculated from their distribution over these channels. The number of combined gas chromatographic-mass spectral channels was limited by using only the 17 most informative mass channels and combining these with the gas chromatographic channels as determined previously. The binary information content of the distribution of the compound occurrence-mass spectral peaks over these channels was then calculated. Since no corrections were made for correlation or for residual errors, the calculated information contents should represent upper limits to the true values.

The class modeling and object classification were performed with the SIMCA 3B software. The data were preprocessed by binary encoding of only those 17 mass channels selected after calculation of the Shannon information content. All data were class scaled. The training sets for each of the classes were selected after investigating the inherent structure of the data set. All objects in a given class were used for the training sets, except for bromobenzene which was only used in the total aromatic training set. The numbers of compounds used in the training sets for the classes were 23 for total aromatics; 51 for total alkanes and alkenes; 8 for chloroaromatics; 14 for nonhaloaromatics; 21 for bromo-alkanes and -alkenes; and 30 for chloroalkanes and -alkenes. Cross validation was used to determine the minimum number of statistically significant components for each class model. Variables with modeling powers less than 0.18 for the principal component models were deleted in the initial stages of the refinement of the class models.

RESULTS AND DISCUSSION

Pattern Recognition of Classes from Mass Spectra

The binary encoded mass spectra of the 78 compounds yielded information contents per channel of 0.10 to 1.00 bit with the higher information at mass channels below 107. Of the channels from 107 and below, 70% had 0.5 bit or greater information. Of the channels greater than 107, only 7.4% had 0.5 bit or greater information content. The information content for the full intensity spectra of the 78 compounds ranged from 0.10 to 2.29 bit with no channels above 107 exceeding 1.02 bit. In general the information contents of the binary and full intensity mass spectra show a linear correlation up to a binary information content of ca. 0.9 bit. Even in the case of the higher information channels there is good qualitative agreement as shown in Table II where the 17 mass channels with the highest binary spectral information content are compared with those from the full intensity spectra.

Table II also contains the information content of the same masses evaluated for a set of 9600 binary mass spectra (5). Comparison of the information content of these 17 mass channels for the binary spectra of the 78 compounds with those found for the set of 9600 compounds also shows a good correlation. The 17 highest information mass channels for the set of 78 compounds yielded 0.80 to 1.00 bit with a median of 0.92 bit, while the same mass channels with the set of 9600 compounds gave 0.62 to 1.00 bit with a median of 0.91 bit. These channels have the highest information content, i.e. greater than 0.80 bit, for the binary encoded spectra with the exception of channel 39, which is not listed. This latter mass channel was not used since it is one of the most frequently occurring peaks in mass spectrometry (6). Thus it is clear that this set

of 17 most "informative" mass channels contains very much information not only with regard to the set of 78 compounds but also with regard to the much larger set of 9600 compounds.

It is well known that the information in a complete mass spectrum is highly redundant and that compressed binary encoded spectra retain a large portion of the information present in the complete spectrum (5,7,8). To solve the present pattern recognition problem on a microcomputer, the number of mass channels considered was reduced from 151 to the 17 channels listed in Table II. This set of binary encoded, 17 mass channel spectra, which will be designated hereafter as the compressed spectra, was used as the trial set of variables in the pattern recognition classification of the 78 compounds. The use of the compressed set of binary spectra caused nine pairs and one trio of compounds to have identical mass spectral representations.

The inherent structure of the data should determine the nature of the classes and not some preconceived scheme based on chemical training or intuition. For a preliminary overview of the data a two dimensional principal component plot of the 17 mass, full intensity data was constructed. Only 49 of the 78 compounds were considered in this plot due to limitations of the SIMCA program used. However, the compounds selected included representatives of all apparent chemical classes in the set. There was no class separation of any type visually apparent in the resulting plot. This result was not unexpected since it has been pointed out in previous studies (9,10) that untransformed mass spectral data should not be used for pattern recognition studies. An examination of the binary encoded, 17 mass data for the same 49 compounds in a two dimensional principal component plot gave the results shown in Figure 1. It is apparent from this plot that there is some basic separation of the compounds into at least three classes and probably four. In the lower right corner is a bromo substituted group of alkanes/alkenes including bromochloro substituted alkanes/alkenes. In the lower left corner is a group of nonhalogenated aromatic compounds. At the top center is a group of chloroalkanes/alkenes. Lying between and partially overlapping the latter two groups is a group of chloroaromatics.

It is clear that the compressed set of data contains enough information for useful classification of the 78 compounds. We have chosen to use two general classes: aromatics with 23 compounds and alkanes/alkenes with 51 compounds. For further detailed classification four subclasses were used: nonhalogenated aromatics with 14 compounds; chloroaromatics with 8 compounds; chloroalkanes/alkenes with 30 compounds; and bromo-alkanes/alkenes with 21 compounds. The seven alkanes/alkenes with both chloro and bromo substituents were found during initial calculations to fit with the bromo alkaenes. The additional four compounds, the three ethers and one epoxide, were withheld from the training sets and used as test objects.

With the use of the variable modeling power and the compressed set of 17 mass spectra, it was found that only five masses were required for each refined class model except for the total alkanes/alkenes class. In the total alkanes/alkenes case 12 masses were required. The use of cross validation resulted in the refined models having only one principal component per model except for the total alkanes/alkenes model which had two. The model parameters including loadings are given in Reference 11. The unexplained variance ranged from a low of 19% for the chloroaromatic model to a high of 47% for the chloroalkane/alkene model. The residual standard deviations, which are a measure of the spatial extent of the models, ranged from 0.49 for the chloroaromatic model to 0.77 for the chloroalkane/alkene model.

After the class models were determined, the entire set of compressed spectra of the 78 compounds was run through all of the six models to determine classification accuracy. Compounds were considered as members of the class to which they were closest. In the case of the two major classes, aromatics and alkanes/alkenes; the training set accuracy was 91 and 98%, respectively, with an overall accuracy of 96%. Only 1 and 3% of the 74 compounds were incorrectly classified as members of the two classes. All of the four test set compounds, which do not properly belong to any of the classes, were classified as alkanes/alkenes, which is correct when interpreted as "not aromatic".

In the case of the four subclasses: nonhaloaromatics, chloroaromatics, bromoalkanes/alkenes, and chloroalkanes/alkenes; the training set accuracy was 79, 87, 86, and 80%, respectively. The overall accuracy was 82%. The number of compounds misidentified as members of the four classes were 3, 4, 3, and 8%, respectively. In this case bromobenzene was added to the four test set compounds. The two chlorosubstituted alkane/alkene type compounds were correctly classified. Bromobenzene was classified as a chloroaromatic which is the most suitable of the four possible classes. The two cyclic alkane ethers were classified as bromoalkane/alkenes, which if regarded as "alkane/alkene, but not chloroalkane/alkene", is correct within the available models.

It was also possible to construct a class model for chlorinated alkane/alkenes based upon the number of chloro groups present in a compound. A single principal component model using the five masses - 63, 65, 75, 77, 91- was found to be statistically valid for the class of mono- and dichloroalkane/alkenes, which did not include compounds substituted with both chloro and bromo groups. This model gave an unexplained variance of 42% for the 16 training set compounds with a residual standard deviation of 0.73. The loadings of the principal component were ca. 0.4-0.5 with mass 63 and 65 having negative loadings. The training set accuracy was only 62%, but the overall accuracy for 78 compounds was 91%. Only two compounds, both monochloromonobromoalkanes/alkenes, were misidentified as members of this class.

Information Content of Analytical Methods

The relative retention times of the 78 compounds on the SE-30 column and their reproducibility resulted in 45 discrete gas chromatography channels for this problem. No compounds were eluted in ten of these channels, so there were only 35 nonzero channels to consider. Since there were less than one-half as many channels as compounds, complete separation (one compound per channel) was not achieved with this column. Of the 35 channels 8.5% contained five compound peaks in each; 5.7% had four compound peaks; 17% had three compound peaks; 37% had two compound peaks; and 31% had only one compound peak. The zero-filled channels contribute no information content. The maximum information content of the 35 gas chromatography channels was 6.41 bits, which was obtained by summing the individual channel information contents.

In the case of the mass spectral channels only the 17 channels used in the pattern recognition study were considered in order to keep the calculations at a manageable size. The largest correlation coefficient among these mass channels was 0.75 between masses 81 and 93, and only four pairs of masses had correlation coefficients greater than 0.5. The maximum information content of these 17 mass spectral channels was 15.4 bits.

One would expect the number of channels available for the combined gas chromatography-mass spectral method to be the product of the two individual methods, i.e. 45 x 17 = 765. This is illustrated in Figure 2, where N =45 and M =17. The shaded areas are zero-filled channels. However, due to the ten zero-filled gas chromatography channels there is a general loss of 170 combined channels, leaving 595 channels. The maximum information content of the combined method can not exceed the sum of the information contents of the individual methods and can not be less than the smaller of the two methods. Therefore the maximum information content of the combined methods is less than 21.8 bits and, with much less certainty, greater than 6.4 bits.

The number of channels, the information contents for the three methods, and the maximum and actual number of compounds distinguishable are summarized in Table III. The maximum number of compounds distinguishable was calculated by using the maximum bits of information as a power of two to obtain the maximum number of different spectral representations. The actual number of compounds distinguished is the average number which is found to be different in a given method. Even though the theoretical maximum number of compounds which can be distinguished is 78 for gas chromatography to ca. four million for gas chromatography; the actual number is much less and only amounts to 76 for the combined gas chromatography-mass spectrometry method. Correlation and uncertainties remaining after analysis account for much of the difference, but some of the difference is due to excess capacity which is not relevant to this particular analysis problem.

CONCLUSIONS

Shannon information theory has been used in this study for selection of relevant variables in SIMCA pattern recognition studies of mass spectral data and as a quantitative measure of maximum information content of analytical methods for a specific problem. In the pattern recognition study of the 78 toxic compounds 151 different mass channels were compressed into 16 which were actually used in the different class models. The major classes found were nonhaloaromatic, chloroaromatic, bromoalkane/alkene, and chloroalkane/alkene compounds. A subclass of alkanes/alkenes containing one or two chlorosubstituents was also found. In all models except the all alkane/alkene model only one principal component containing five masses was found to be statistically significant. The accuracy was 96% for alkane/alkene vs. aromatic classification and 82% for classification into the four groups.

SIMCA pattern recognition is an empirical classification technique and does not require particular spectroscopic expertise for its application. Its successful application to the present problem on a microcomputer with limited memory should demonstrate to chemists the availability and power of the technique and the large amount of useful information in the low resolution mass spectra.

The numbers of channels and information content for a given method should not be confused with those relevant to the solution of a particular problem. As has been shown here the amount of maximum information for a given analytical method may be very large, but all of the capacity of the method may not be applicable to a specific analytical problem.

REFERENCES

1. P. Cleij, A. Dijkstra, Z. Anal. Chem., 298: 97 (1979).

2. S. Wold, C. Albano, W. J. Dunn III, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg, W., M. Sjostrom, in B. R Kowalski, Ed., "Chemometrics, Mathematics and Statistics in Chemistry", D. Reidel Publishing Co., Boston, 1984, pp. 17-96.

3. S. Wold, Pattern Recognit., 8: 127 (1976).

4. S. Wold, Technometrics, 20: 127 (1978).

5. G. van Marlen, A. Dijkstra, Anal. Chem., 48: 595 (1976).

6. G. M. Pesyna, F. W. McLafferty, R. Venkataraghavan, H. E. Dayringer, Anal. Chem., 47: 1161 (1975).

7. S. L. Grotch, Anal. Chem., 42: 1214 (1970).

8. J. B. Justice, T. L. Isenhour, Anal. Chem., 46: 223 (1974).

9. J. R. McGill, B. R. Kowalski, J. Chem. Inf. Comput. Sci., 18: 52 (1978).

10. S. Wold, H. J. Christie, Anal. Chim. Acta, 165: 51 (1984).

11. D. R. Scott, Anal. Chem., 58: 881 (1986).

TABLE I. COMPOUNDS INCLUDED IN THIS STUDY

1. p-xylene
2. 1,3,5-trimethylbenzene
3. isopropylbenzene
4. n-butylbenzene
5. 1-methyl-4-isopropylbenzene
6. o-dichlorobenzene
7. p-dichlorobenzene
8. 1-chloro-2-methylbenzene
9. 1-chloro-4-methylbenzene
10. p-chlorostyrene
11. 1,1-dichloroethane
12. 1,1,1,2-tetrachloroethane
13. 1,2,3-trichloropropane
14. 3-chloropropene
15. 2-chlorobutane
16. 1,3-dichlorobutane
17. 1,4-dichlorobutane
18. 1,4-dichloro-2-butene (cis)
19. 3,4-dichlorobutene
20. 1,4-dioxane
21. 1-chloro-2,3-epoxypropane
22. 2-chloroethoxyethene
23. acetophenone
24. benzonitrile
25. benzene
26. toluene
27. o-xylene
28. m-xylene
29. ethylbenzene
30. styrene
31. chlorobenzene
32. bromobenzene
33. m-dichlorobenzene
34. 1-chloro-3-methylbenzene
35. chloroform
36. carbon tetrachloride
37. bromochloromethane
38. bromotrichloromethane
39. dibromomethane
40. bromoform
41. 1,2-dichloroethane
42. 1,1,1-trichloroethane
43. 1,1,2-trichloroethane
44. 1,1,2,2-tetrachloroethane
45. pentachloroethane
46. 1,1-dichloroethene
47. trichloroethene
48. tetrachloroethene
49. bromoethane
50. 1,2-dibromoethane
51. 1-chloropropane
52. 2-chloropropane
53. 1,2-dichloropropane
54. 1,3-dichloropropane
55. 1-bromo-3-chloropropane
56. 1,2-dibromopropane
57. 2,3-dichlorobutane
58. tetrahydrofuran
59. benzaldehyde
60. 1-bromo-1-chloroethane
61. 2,2-dibromopropane
62. 2-bromopropene
63. 2-bromopropane
64. 3-bromopropene
65. 1-bromopropane
66. 1-chlorobutane
67. 1-bromo-2-chloroethane
68. bromodichloromethane
69. 1-bromobutane
70. 2,2-dichlorobutane
71. dibromochloromethane
72. 1,1,2-trichloropropane
73. 1,3-dibromopropane
74. 1,1,1,2-tetrachloropropane
75. 1,2,2,3-tetrachloropropane
76. 1,3-dibromobutane
77. 1,1,2,3-tetrachloropropane
78. 1,4-dibromobutane

TABLE II. 17 KEY MASSES WITH HIGHEST INFORMATION CONTENT FOR 78 TOXIC COMPOUNDS

MASS	INFORMATION CONTENT			MASS	INFORMATION CONTENT		
	BINARY[a]	FULL[b] (BITS)	9600[c]		BINARY	FULL (BITS)	9600
41	0.92	1.92	0.90	77	0.94	1.75	1.00
49	0.98	1.68	0.62	78	0.84	1.31	0.93
50	0.84	1.18	0.98	79	0.96	1.27	0.97
51	0.99	1.78	1.00	81	0.80	1.02	0.91
61	0.82	1.36	0.75	91	0.92	1.47	0.96
62	0.95	1.61	0.81	93	0.87	1.22	0.83
63	1.00	1.95	0.97	95	0.90	1.39	0.83
65	0.92	1.35	0.98	107	0.82	1.08	0.74
75	0.89	1.42	0.89				

a. Binary spectra of 78 compounds.
b. Full intensity spectra of 78 compounds.
c. Binary spectra of 9600 compounds, Reference 5.

TABLE III. GC, MS, GC-MS INFORMATION CONTENT FOR 78 COMPOUNDS USING 17 KEY MASSES

METHOD	CHANNELS		MAXIMUM I (BIT)	CMPDS DISTINGUISHABLE	
	TOTAL	NONZERO		MAXIMUM[a]	ACTUAL[b]
GC	45	35	6.4	78	27
MS	17	17	15.4	43000	60
GC-MS	765	595	< 21.8	< 3700000	76

a. The number obtained by raising two to the power of maximum I.
b. The number of different compounds actually distinguished by the method.

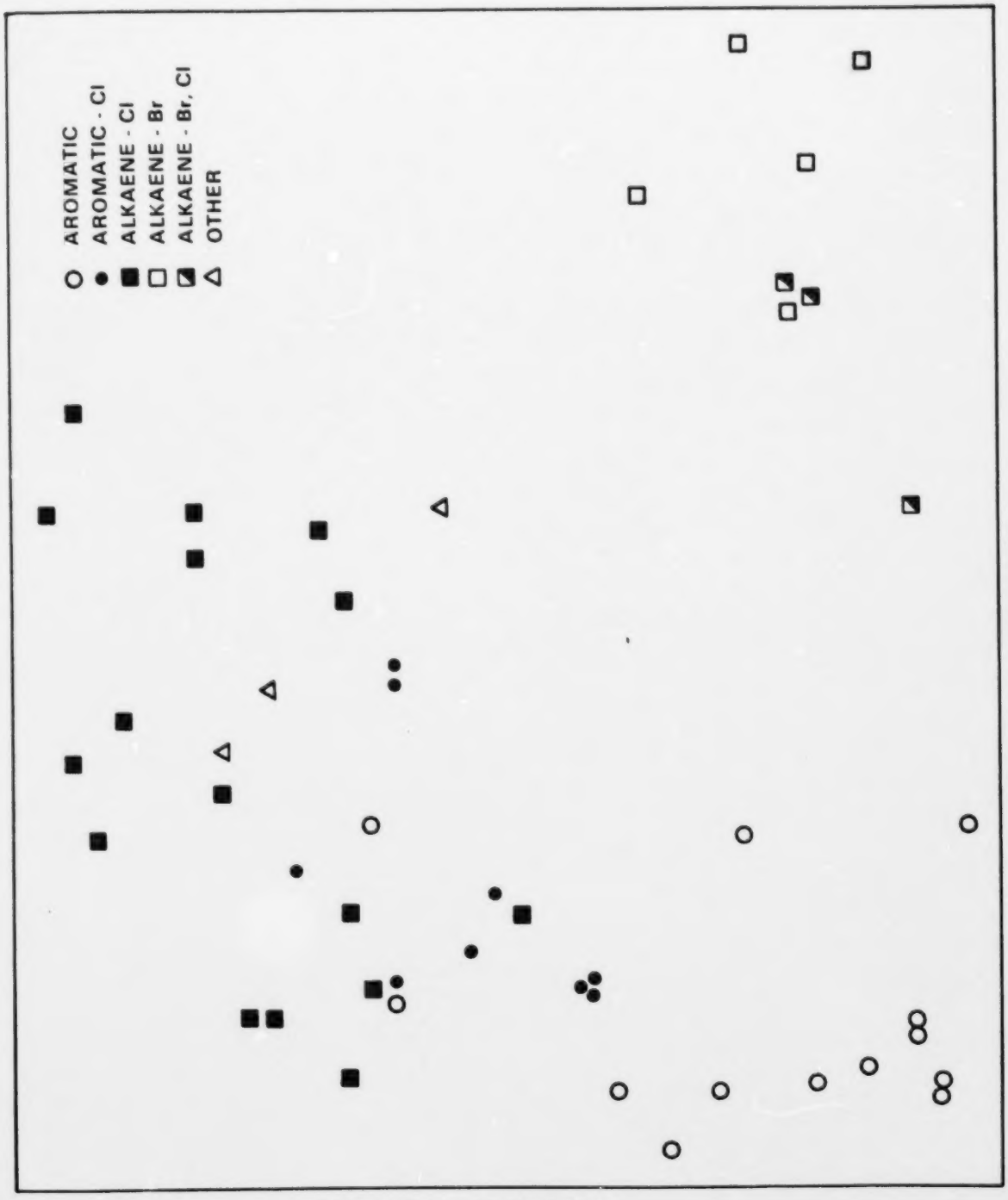


Figure 1. Principal Component Plot of 17 Key Mass Spectra of 49 Compounds.

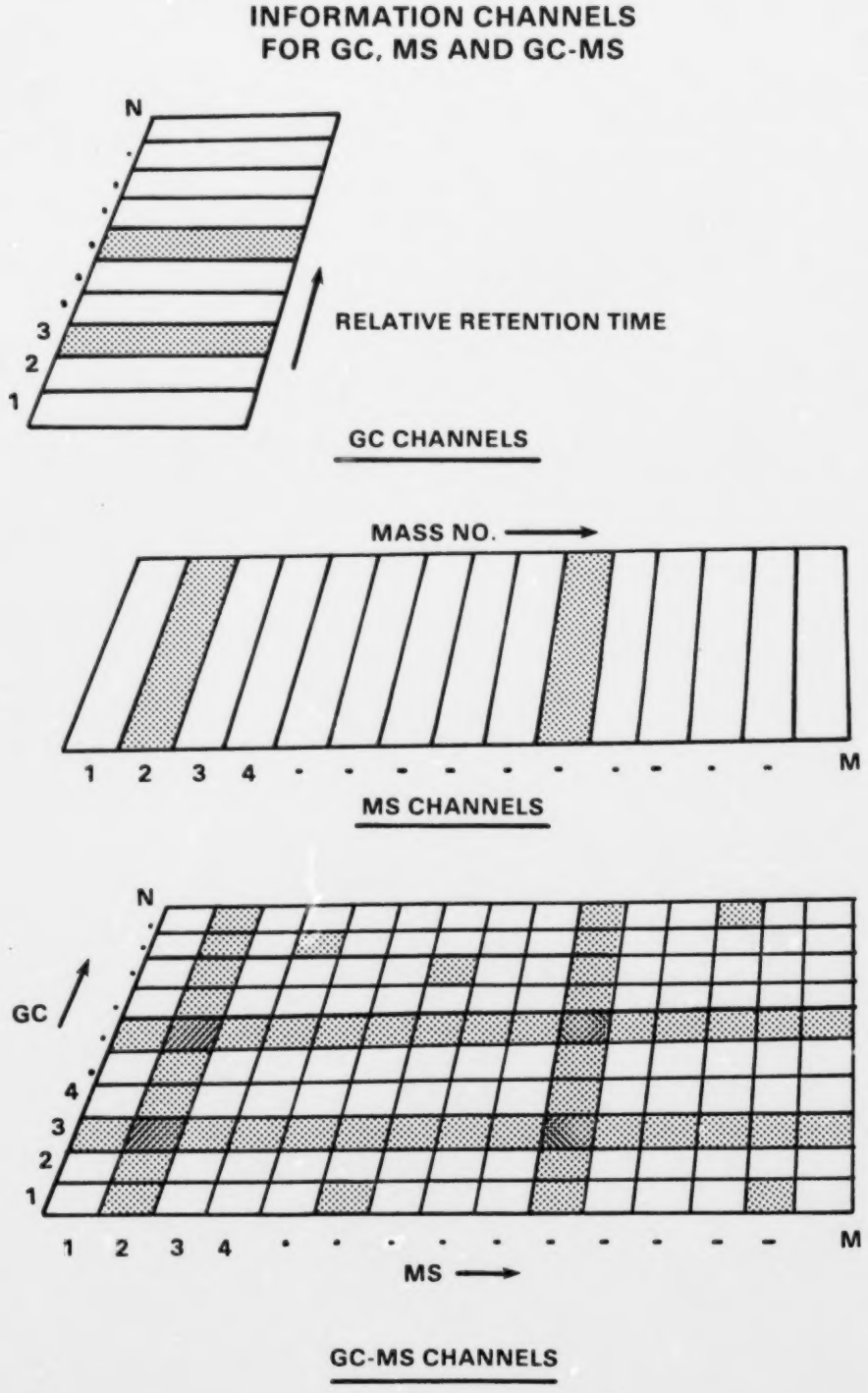


Figure 2. Combined GC-MS Channels from GC and MS Channels.